ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

BÙI ĐỨC VIỆT

PHÂN CỤM DỮ LIỆU CHO NHẬN DẠNG ẢNH SỬ DỤNG MẠNG NƠRON

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN, NĂM 2012

Số hóa bởi Trung tâm Học liệu – Đại học Thái Nguyên *http://www.lrc-tnu.edu.vn*

LỜI CẢM ƠN

Trước tiên em gửi lời cảm ơn chân thành sâu sắc tới các thầy cô giáo ở Viện Công nghệ thông tin Việt Nam, các thầy cô trong trường Đại học Công nghệ thông tin & Truyền thông - Đại học Thái Nguyên đã tận tình truyền đạt, giảng dạy cho em những kiến thức, kinh nghiện quý báu trong suốt thời gian qua.

Đặc biệt em xin gửi lời cảm ơn đến PGS.TS Lê Bá Dũng đã tận tình giúp đỡ, trực tiếp chỉ bảo em trong suốt thời gian làm luận văn. Trong thời gian làm việc với Thầy, em không những tiếp thu thên nhiều kiến thức bổ ích mà còn học được tinh thần làm việc, thái độ nghiên cứu khoa học nghiêm túc, hiệu quả. Đây là những điều rất cần thiết cho em trong quá trình học tập và công tác.

Sau cùng xin gửi lời cảm ơn chân thành tới gia đình, bạn bè đã động viên, đóng góp ý kiến và giúp đỡ trong quá trình học tâp, nghiên cứu và hoàn thành đề tài này.

Thái Nguyên, tháng 10 năm 2012

Học viên

Bùi Đức Việt

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC TỪ VIẾT TẮT

CSDL	Cơ sở dữ liệu
PCDL	Phân cụm dữ liệu
KPDL	Khai phá dữ liệu
BNU	Phần tử nơron chiến thắng
MLP	MultiLayer Perception
BAM	Bidirectional Associative Memory
SOM	Self Organizing Map
VQ	Vector Quantization
LVQ	Learning Vector Quantization
MST	Minimal Spanning Tree

DANH MỤC CÁC HÌNH VẼ

LỜI NÓI ĐẦU

Sự phát triển của công nghệ thông tin và việc ứng dụng công nghệ thông tin trong nhiều lĩnh vực của đời sống, kinh tế xã hội nhiều năm qua cũng đồng nghĩa với lượng dữ liệu đã được các cơ quan thu thập và lưu trữ ngày một tích lũy nhiều lên. Nếu cho rằng, điện tử và truyền thông chính là bản chất của khoa học điện tử thì dữ liệu, thông tin, tri thức là tiêu điểm của một lĩnh vực mới để nghiên cứu và ứng dụng, đó là khám phá tri thức và khai phá dữ liệu.

Thông thường, chúng ta coi dữ liệu là một chuỗi các bits, hoặc các số và các ký hiệu hay các đối tượng với một ý nghĩa nào đó khi gửi cho một chương trình dưới một dạng nhất định. Các bít thường được sử dụng để đo thông tin, và xem nó như là dữ liệu đã loại bỏ phần tử dư thừa, lặp lại và rút gọn tới mức tối thiểu để đặc trưng một cách cơ bản cho dữ liệu. Tri thức được xem như là những thông tin tích hợp, bao gồm các sự kiện và các mối quan hệ giữa chúng đã được nhận thức, khám phá hoặc nghiên cứu. Nói cách khác, tri thức có thể được coi là dữ liệu ở mức độ cao của sự trừu tượng và tổng quát.

Khám phá tri thức hay phát hiện tri thức trong CSDL là quy trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: phân tích, tổng hợp, hợp thức, khả ích và có thể hiểu được.

Khai phá dữ liệu là một bước trong quá trình khám phá tri thức, gồm các thuật toán khai thác dữ liệu chuyên dùng dưới một số quy định về hiệu quả tính toán chấp nhận được để tìm ra các mẫu hoặc các mô hình trong dữ liệu. Nói cách khác, mục tiêu của khai phá dữ liệu là tìm kiếm các mẫu hoặc các mô hình tồn tại trong CSDL nhưng ẩn trong khối lượng lớn dữ liệu.

Phân cụm dữ liệu (PCDL) là quá trình nhóm một tập các đối tượng tương tự nhau trong tập dữ liệu vào các cụm sao cho các đối tượng thuộc cùng một cụm là tương đồng còn các đối tượng thuộc các cụm khác nhau sẽ không tương đồng. Phân cụm dữ

liệu là một ví dụ của phương pháp học không có thầy. Không giống như phân lớp dữ liệu, phân cụm dữ liệu không đòi hỏi phải định nghĩa trước các mẫu dữ liệu huấn luyện. Vì thế, có thể coi phân cụm dữ liệu là một cách học bằng quan sát, trong khi phân lớp dữ liệu là học bằng ví dụ…

Hiện nay, các phương pháp phân cụm trên đã và đang được phát triển và áp dụng nhiều trong các lĩnh vực khác nhau và đã có một số nhánh nghiên cứu được phát triển trên cơ sở của các phương pháp đó như:

Phân cụm thống kê: Dựa trên các khái niệm phân tích hệ thống, nhánh nghiên cứu này sử dụng các độ đo tương tự để phân hoạch các đối tượng, nhưng chúng chỉ áp dụng cho các dữ liệu có thuộc tính số.

Phân cụm khái niệm: Kỹ thuật này được phát triển áp dụng cho dữ liệu hạng mục, chúng phân cụm các đối tượng theo các khái niệm mà chúng xử lí.

Phân cụm mờ: Sử đụng kỹ thuật mờ để PCDL. Các thuật toán thuộc loại này chỉ ra lược đồ phân cụm thích hợp với tất cả các hoạt động đời sống hàng ngày, chúng chỉ xử các dữ liệu không chắc chắn.

Luận văn gồm có 4 chương:

Chương 1: Giới thiệu về khai phá dữ liệu

Chương 2: Tổng quan về phân cụm dữ liệu

Chương 3: Ứng dụng mạng Nơron Kohonen cho phân cụm dữ liệu

Chương 4: cài đặt thử nghiệm

Luận văn đã trình bày một số vấn đề về phân cụm - một trong những kỹ thuật cơ bản để khai phá dữ liệu và ứng dụng phân cụm cho nhận dạng ảnh sử dụng mạng nơron. Đây là hướng nghiên cứu có triển vọng chỉ ra những sơ lược trong việc hiểu và khai thác CSDL khổng lồ, khám phá thông tin hữu ích ẩn trong dữ liệu; hiểu được ý nghĩa thực tế của dữ liệu và ứng dụng vào bài toán cụ thể.